

Aplicações Informáticas na Biomedicina

7.^a Aula Prática Laboratorial

Mestrado Integrado em Engenharia Informática

Ano Letivo 2019/2020

Marisa Esteves

6 de Novembro de 2019



Universidade do Minho

Plano de Aula

1. Contextualização e demonstração do Talend;
2. Resolução da 4.^a ficha prática laboratorial pelos alunos em grupo;
3. Correção da ficha com os alunos.

Processo ETL

Definição

O processo ETL (*Extract, Transform, Load*) é um conjunto de processos que inclui a extração de dados de fontes de informação internas e externas, podendo estar em diferentes formatos, a transformação dos dados de acordo com as necessidades da organização e, finalmente, o carregamento dos mesmos numa estrutura de dados, como por exemplo um data mart ou um data warehouse.

Processo ETL

Definição

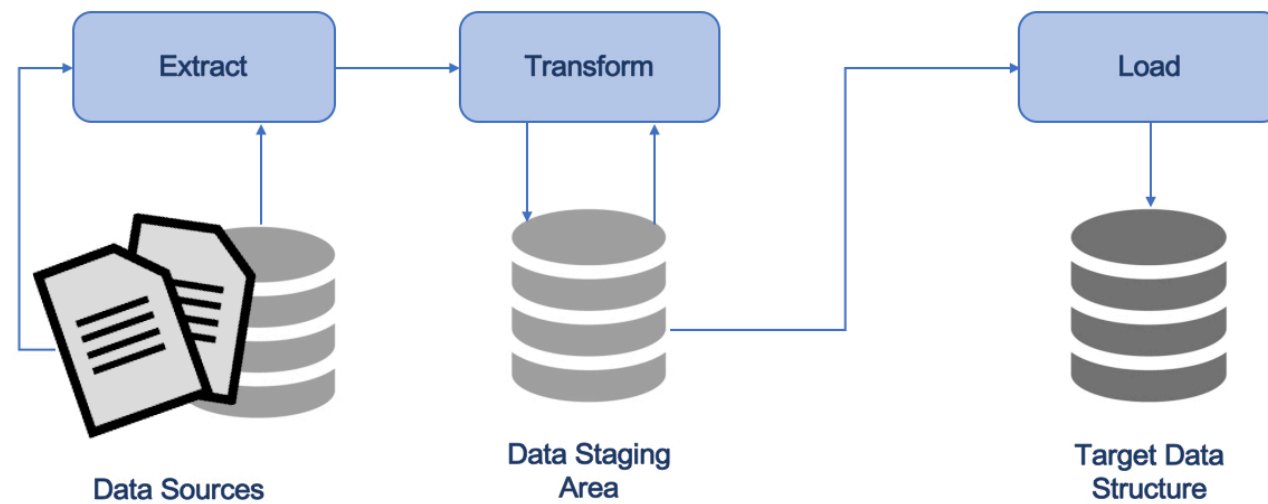


Figura 1 – Esquema do processo ETL.

Processo ETL

Porquê?

Os dados estão espalhados
por diferentes localizações

Os dados estão
armazenados em diferentes
tipos de formato

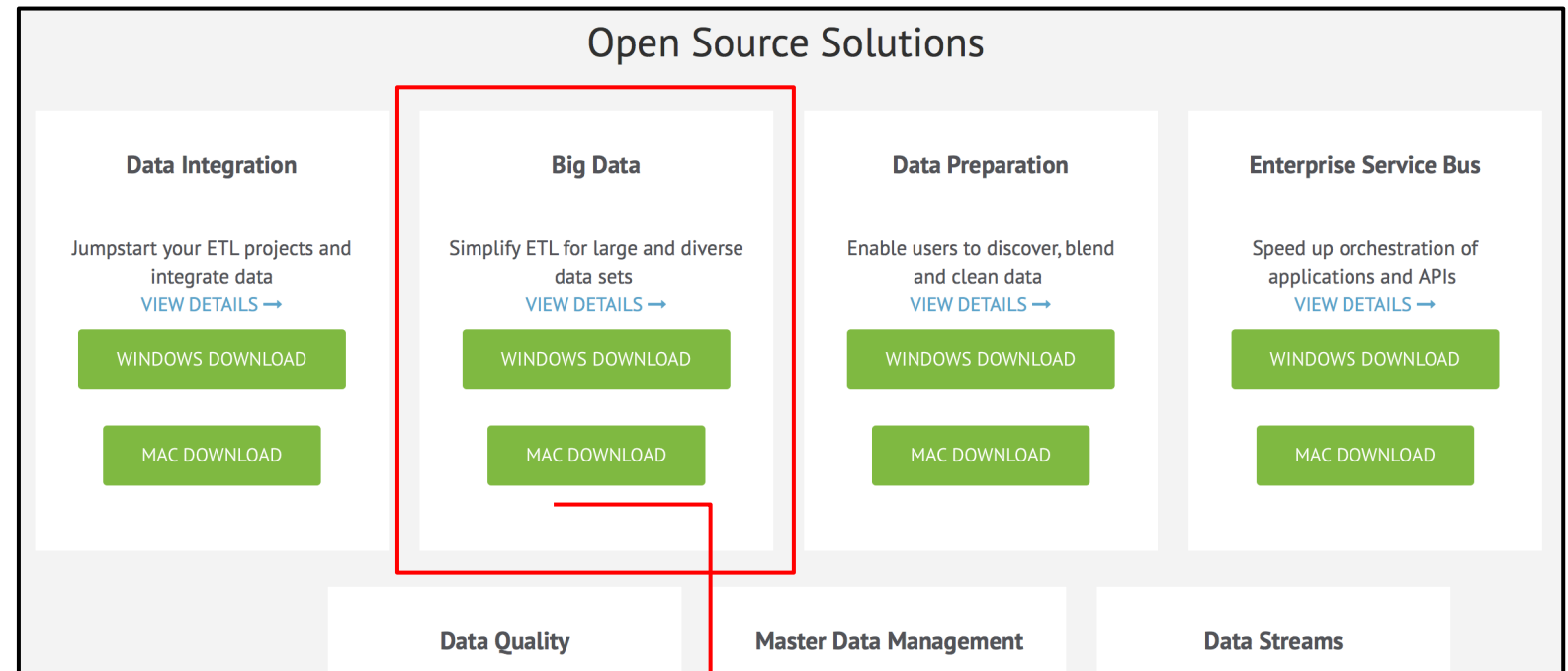
O volume de dados continua
a aumentar

Os dados podem estar
estruturados, semi-
estruturados ou não
estruturados

Instalação

Talend

<https://www.talend.com/products/talend-open-studio/>



+ third party libraries (jars)

Talend

Contextualização

Ferramentas ETL – São ferramentas que combinam as três fases do processo de ETL (*Extract, Transform, Load*) numa única ferramenta.

Fácil de usar

User friendly GUI

Tratamento dos
erros incorporado

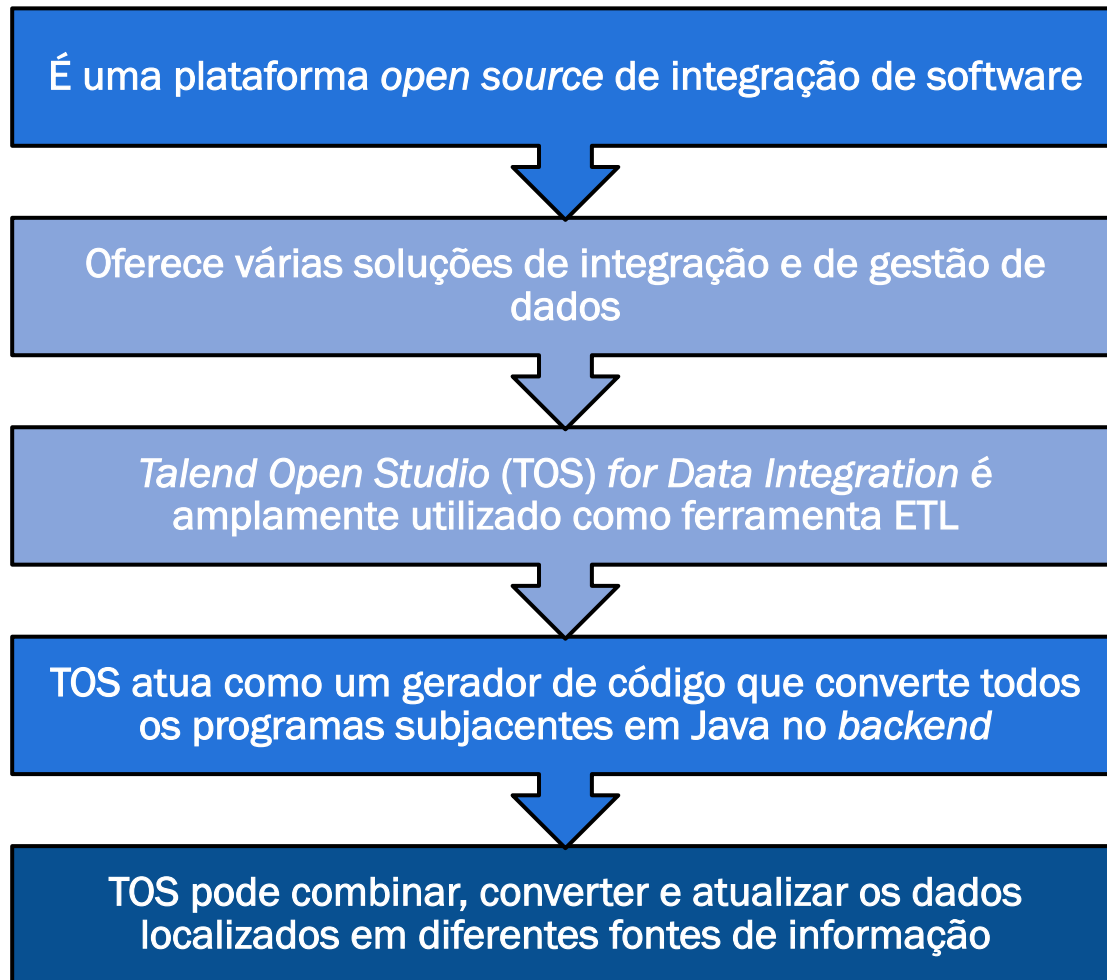
Reduz as despesas

Melhora a gestão
dos dados

Melhora o
desempenho

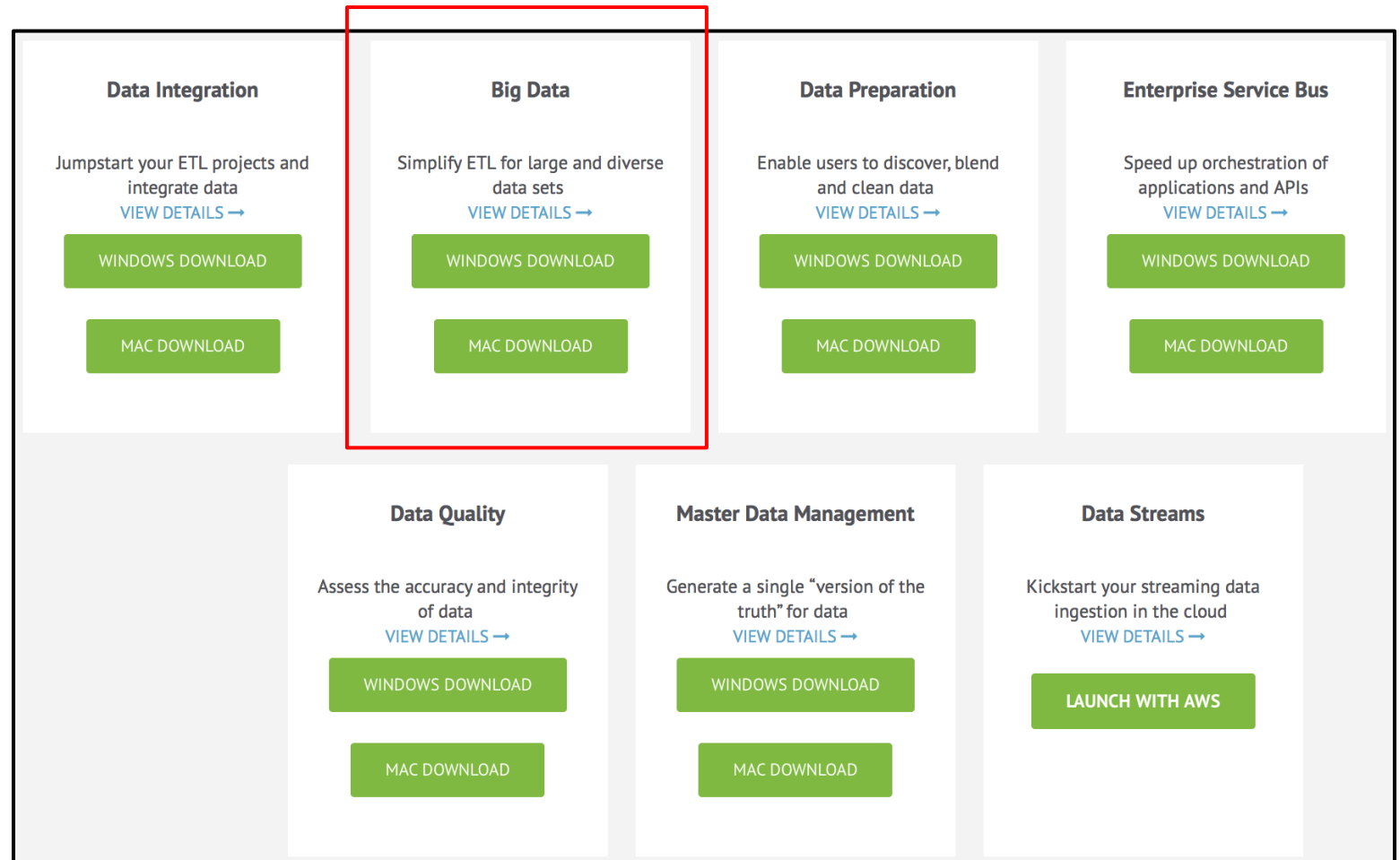
Talend

Definição



Talend

Definição

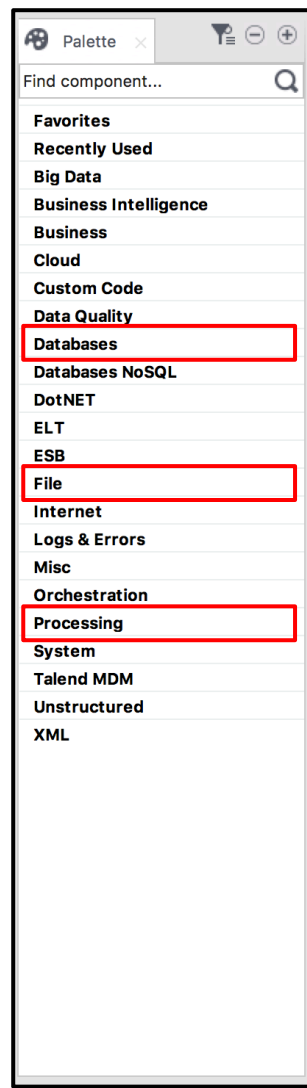


The image shows a grid of seven Talend product categories. The 'Big Data' category is highlighted with a red border. Each category includes a brief description, a 'VIEW DETAILS' link, and download buttons for Windows and Mac. The 'Data Streams' category includes a 'LAUNCH WITH AWS' button instead of download buttons.

Category	Description	VIEW DETAILS →	Windows Download	Mac Download
Data Integration	Jumpstart your ETL projects and integrate data	VIEW DETAILS →	WINDOWS DOWNLOAD	MAC DOWNLOAD
Big Data	Simplify ETL for large and diverse data sets	VIEW DETAILS →	WINDOWS DOWNLOAD	MAC DOWNLOAD
Data Preparation	Enable users to discover, blend and clean data	VIEW DETAILS →	WINDOWS DOWNLOAD	MAC DOWNLOAD
Enterprise Service Bus	Speed up orchestration of applications and APIs	VIEW DETAILS →	WINDOWS DOWNLOAD	MAC DOWNLOAD
Data Quality	Assess the accuracy and integrity of data	VIEW DETAILS →	WINDOWS DOWNLOAD	MAC DOWNLOAD
Master Data Management	Generate a single "version of the truth" for data	VIEW DETAILS →	WINDOWS DOWNLOAD	MAC DOWNLOAD
Data Streams	Kickstart your streaming data ingestion in the cloud	VIEW DETAILS →	LAUNCH WITH AWS	

Talend

Componentes



A ferramenta inclui mais de 900 componentes e conectores divididos em famílias de componentes. As mais utilizadas são as seguintes:

File Components

- tFileInputDelimited;
- tFileInputExcel;
- tFileOutputDelimited;
- tFileOutputExcel;
- tFileList;
- tFileExists;
- tFileCopy.

Processing Components

- tMap;
- tJoin;
- tFilterRow;
- tSortRow.

Databases Components

- tDBConnection;
- tDBInput;
- tDBRow;
- tDBCommit;
- tDBOutput.

Demonstração

Talend

The screenshot displays the Talend Studio interface for a job named 'demo_job1 1.1'. The main workspace shows a job design with the following components and flow:

- mysql_localhost_talend** (Database Connection) connects to **OnSubjobOk** (Job Trigger).
- OnSubjobOk** triggers **lista_espera_blo** (File Input Delimited).
- lista_espera_blo** feeds into **tMap_1** (Map).
- tMap_1** outputs to **output (Main)**, which then feeds into **tDBOutput_3** (Database Output).
- tDBOutput_3** outputs to **row2 (Main)**.
- row2 (Main)** feeds into **tFileOutputExcel_1** (File Output Excel).
- tFileOutputExcel_1** outputs to **row4 (Main)**.
- row4 (Main)** feeds into **tSortRow_1** (Sort Row).
- tSortRow_1** outputs to **row3 (Filter order:1)**.
- row3 (Filter order:1)** feeds into **tFilterRow_1** (Filter Row).
- tFilterRow_1** outputs to **row3 (Reject order:2)**.
- row3 (Reject order:2)** feeds into **tFileOutputExcel_2** (File Output Excel).

The interface includes a **Repository** pane on the left showing the project structure, a **Palette** on the right for component selection, and an **Execution** pane at the bottom for running and debugging the job.

Resolução da 4.ª Ficha Prática Laboratorial

1 Criação e Execução da Primeira Job no Talend

O ficheiro disponibilizado juntamente com esta ficha prática laboratorial, nomeadamente “world_happiness_2016.csv”, contém os dados reais relativos à classificação da felicidade por país em 2016 de acordo com diversas variáveis.

O *dataset* é constituído por 13 diferentes colunas, nomeadamente *country*, *region*, *happiness_rank*, *happiness_score* (de 0 a 10), *lower_confidence_interval*, *upper_confidence_interval*, *economy*, *family*, *health* (esperança de vida), *freedom*, *trust* (corrupção governamental), *generosity* e *dystopia_residual*.

É importante referir que as colunas *economy*, *family*, *health* (esperança de vida), *freedom*, *trust* (corrupção governamental), *generosity* e *dystopia_residual* descrevem em que medida esses fatores contribuíram para avaliar a felicidade em cada país.

Tendo em conta o seu enquadramento teórico e prático sobre o processo de ETL e o Talend, os objetivos principais a serem alcançados com a resolução desta ficha prática laboratorial são a sua ambientação à ferramenta *Talend Open Studio (TOS) for Big Data*, bem como a criação e a execução da sua primeira *job* com as componentes disponibilizadas na mesma.

Resolução da 4.ª Ficha Prática Laboratorial

Com esta ficha prática laboratorial, pretende-se que:

1. Crie um novo *schema* no MySQL Workbench denominado “Ficha4”.
2. Abra o Talend e crie uma *job*, recorrendo às componentes adequadas, que permita, de forma geral:
 - (a) Conectar-se à base de dados criada no passo anterior (*tDBConnection*).
 - (b) Fazer o *import* dos dados no ficheiro denominado “world_happiness_2016.csv” (*tFileInputDelimited*).
 - (c) Fazer o *export* de toda a informação no ficheiro world_happiness_2016.csv para uma nova tabela denominada “world_happiness_2016” no *schema* Ficha4 (*tMap* e *tDBOutput*).
 - (d) Fazer o *export* dos dados na tabela world_happiness_2016 de todos os países em regiões europeias para uma nova tabela denominada “world_happiness_2016_europe” no *schema* Ficha4 (*tFilterRow* e *tDBOutput*).
 - (e) Criar um ficheiro Excel (.xls), nomeadamente world_happiness_2016_europe_good.xls, com os dados relativos aos países europeus que tenham um *happiness_score* superior ou igual a 5 (*tFilterRow* e *tFileOutputExcel*).
3. Corra a *job* criada.